We would like to thank all referees for their appreciation of our results and the useful feedback. Below is our reply.

Reviewer 1: There are (at least) two reasons to justify the averaging over a ball of radius $\gamma > 0$ around x_0. First, Example 3.2 indicates that when $\gamma = 0$, the estimator may be inconsistent. This is equivalent to 1-nearest neighbor is not consistent in general. Second, Example 3.3 shows that we can recover the k-nearest neighbors by choosing γ appropriately. Averaging over a ball around x_0 thus eliminates an undesirable statistical property (inconsistency) and gives us the flexibility to recover k-nearest neighbors for general k.

To improve the transparency of our estimator, we will provide in the revision a description of the worst-case distribution. Just as an adversarial example provides a description on how to perturb a data point, the worst-case distribution provides full information on how to perturb the empirical distribution from the adversary's viewpoint. For our estimator, constructing the worst-case distribution is intriguing and intuitive: it involves sorting the values $v^\star$ defined in equation (5) and then performing a greedy assignment. The construction of this worst-case distribution is done as part of the proof of Theorem 2.3. We agree that this information should be made more explicit to the readers. We will include the worst-case distribution and elaborate more details in the revised version.

Thank you for pointing out the relevant literature. To our understanding, existing robustification of nearest neighbors (and nonparametric classifiers in general) can be divided into two streams: i) global approaches that modify the whole training dataset, e.g., adversarial pruning (arXiv:1706.03922, arXiv:1906.03310, arXiv:2003.06121, etc.), and ii) local approaches that study attack for each data point and find appropriate defense for specific classifiers such as 1-NN (arXiv:1811.00525, arXiv:1906.03972, etc.).

Compared to the current literature, we believe that our approach is more general in two significant ways: i) we start from a generic min-max estimation problem, and our ideas and methodology are easily applicable to other non-parametric settings, and ii) we allow for perturbations on Y to hedge against label contamination. We will include this discussion.

Reviewer 2: Thank you very much for your feedback. We would like to emphasize that our paper aims to provide a principled approach to robustify nonparametric estimators, our contributions include the proposal of a novel adversarial estimation framework along with theoretical insights.

Regarding the experiment: the MNIST dataset is still the field's standard benchmark dataset to evaluate and compare performance among models (Google Scholar indicates $\sim$591 citations to the MNIST dataset since 2019 alone). To study how robust a deep learning model is subject to (possibly adversarial) distributional shift, the MNIST dataset is also one of the leading benchmarks (arXiv:1906.02530). State-of-the-art research on robustifying nonparametric estimators (arXiv:1706.03922, arXiv:1906.03310, arXiv:2003.06121, arXiv:1811.00525, arXiv:1906.03972, etc.) also focus on simple experimental settings to condense and deliver insights.

Regarding the performance: From Table 2, our estimator outperforms the N-W estimator from 9% ($N = 500$) to 20% ($N = 50$) in terms of accuracy in the MNIST dataset. Further results in the appendix show that we can stochastically dominate other nonparametric approaches in both synthetic (Figure A.2) and MNIST dataset with $p \leq 1$ (Figure A.3).

Reviewer 3: Thank you for your suggestion on the lower bound. Currently we focus on the upper bound because it is the canonical analysis for minimax problems. The probabilistic *lower* bound can be stated in the below result.

Proposition. *Under the same settings of Proposition 3.1, with a probability of at least* $1 - O(N^{-c})$*, we have*

$$\mathbb{E}_{\mathbb{P}}[\ell(Y, \beta^\star)|X \in \mathcal{N}_\gamma(x_0)] \geq - \sup_{\mathbb{Q} \in \mathbb{B}_\rho, \mathbb{Q}(X \in \mathcal{N}_\gamma(x_0)) > 0} \mathbb{E}_{\mathbb{Q}}\big[- \ell(Y, \beta^\star)|X \in \mathcal{N}_\gamma(x_0)\big]$$

To evaluate the supremum on the right-hand side, it suffices to use Theorem 2.3 in the paper with minor changes to the definition of the values $v_i^\star(\beta^\star)$. The proof of this claim follows similar argument as in the proof of Proposition 4.1.

Regarding Example 3.3: Thank you for noticing the linearity condition. We recheck [29, Theorem 2 and Corollary 3] where 'global' consistency is obtained universally in probability. Since we focus on 'local' consistency, only *continuity* of the regression function is required, that means Example 3.3 is valid when $\mathbb{E}_{\mathbb{P}}[Y|X = x]$ is continuous in x. We will correct this condition in the revision and clarify the (stronger) notion of consistency that we have in mind.

Regarding the performance of our estimator: In the experiment, we specifically tune our estimator so that it behaves as a robust k-nearest neighbor estimator. As such, it is more reasonable to compare our estimator versus the vanilla k-nearest neighbor approach. Experiments using both synthetic and the MNIST dataset show that we clearly outperform the k-nearest neighbor, which justifies the benefit of being robust.

Even when comparing our estimator versus the N-W estimator, Table 2 shows that our estimator outperforms the N-W estimator at all sample sizes. The improvement can be as big as 20% ($N = 50$) in terms of accuracy in the MNIST dataset. In the appendix, we also present additional results showing that our estimator can stochastically dominates other nonparametric approach in both the synthetic setting (Figure A.2) and the MNIST dataset with $p \leq 1$ (Figure A.3). Our estimator thus delivers performance gains at reasonable computational overhead.